DOCUMENT RESUME

ED 414 306 TM 027 771

AUTHOR Kimmel, Ernest W.
TITLE Unintended Consequences or Testing the Integrity of Teachers and Students.
INSTITUTION Educational Testing Service, Princeton, NJ.
PUB DATE 1997-06-00
NOTE 8p.; Paper presented at the Annual Assessment Conference of the Council of Chief State School Officers (Colorado Springs, CO, June 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; *Achievement Tests; *Cheating; Computer Assisted Testing; Elementary Secondary Education; *Standardized Tests; Teacher Education; Test Validity; *Testing Problems
IDENTIFIERS *Test Security

ABSTRACT

Large-scale testing programs are generally based on the assumptions that the test-takers experience standard conditions for taking the test and that everyone will do his or her own work without having prior knowledge of specific questions. These assumptions are not necessarily true. The ways students and educators use to get around standardizing conditions to gain an advantage are described, and ways to reduce these behaviors are presented. In the first place, there is traditional cheating, by copying, or describing answers, which is enhanced by electronic gadgets or international exchanges of information. Lax security can result in the theft of test booklets. Teachers and other educators can undermine the validity of test by ignoring evidence of student preknowledge of the test or by tacitly colluding with students by allowing access to test materials. The measurement community needs to do a better job of educating educators about the importance of standard conditions. Testing agencies or programs should audit some testing sites to determine the existence of standard conditions, and new techniques of test administration, including computer adaptive tests, must be developed to improve test security. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

UNINTENDED CONSEQUENCES
OR
TESTING THE INTEGRITY OF TEACHERS AND STUDENTS

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Ernest Kimmel
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Ernest W. Kimmel
Educational Testing Service

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

National Conference on Large Scale Assessment
Colorado Springs, CO
June 1997

ED 414 306

Virtually all large-scale testing or assessment programs are based on the premise that the test-takers experience standard conditions for taking the test. It is assumed that everyone will experience the same timing, the same environment for taking the test, and the same process for scoring responses. It is further assumed that everyone does his or her own work on the assessment and has no prior knowledge of specific questions.

While we might profitably discuss how the disregard of any of these conditions affects the validity of the results, I want to focus on those last two assumptions and the ways in which they are abused, especially in testing programs with important consequences for individual students or particularly for teachers and administrators when the results are used for accountability purposes, such as part of a state report card.

TM027771

Almost half a century ago, in the first edition of *Educational Measurement*,[1] Arthur Traxler argued that valid results from testing depended upon accurate administration and scoring. He went on to complain that "In view of their crucial importance in the whole chain of events from the conception of the test to the use of the scores . . . it seems highly unfortunate that the giving and scoring of tests are frequently treated very casually . . . ." In his discussion of the various operational factors that might undermine the validity of a test, he included "purposeful copying" as a subsidiary source of error. Later in the chapter, he addressed the security issue by suggesting that the testing room not be overcrowded and that all

BEST COPY AVAILABLE

students be tested at the same time. He also observed that ". . . some teachers who are testing their own classes may be so eager for them to do well that they will yield to temptation to offer a few indirect suggestions which will help the pupils obtain higher scores."

Our large-scale testing programs -- whether state, national, or international -- assume that the test administrators and test takers play by the same rules where-ever they are. Even though Traxler, writing in the late '40s, recognized and deplored certain threats to test security, the basic assumption of standardized test administrations worked reasonably well. Although individual test takers violated the rules, they were usually "caught" by the teacher or other proctor or test administrator. Such seeming incidental cheating did not seem to seriously threaten the validity of the interpretations being made of the test results.

However, because test results are increasingly seen as more credible measures of student achievement than other information, critical decisions are being based solely or primarily on test scores. For example, with the Hopwood decision in Texas, admissions decisions to the public universities had to be made this year primarily on the basis of test scores. Elsewhere, teachers are being evaluated on the test scores earned by their students. Concomitantly, there is growing evidence that the pressure of such high stakes decisions based on the test scores leads some test takers and teachers to violate the implicit social contract of obtaining test scores under standard conditions.

I will describe the variety of ways that students and educators use to "beat" the standardizing conditions to gain an advantage and then suggest a number of ideas that might be pursued to reduce those behaviors that undermine our ability to interpret scores as deriving from the same conditions.

The practice of ignoring or violating the basic ground rules of test administration, or put more kindly, the casualness with which instructions are implemented, has serious implications for the proposed Voluntary National Tests. If the national tests are used for any serious decisions about

students or schools, there will be enormous pressures on the security of the tests. With a highly distributed structure for administering them, it will be hard to determine whether the standard conditions prescribed by the developers are being provided.

What kinds of things are happening? Let me describe some of the situations we've encountered at ETS. Similar situations undoubtedly occur with other testing and assessment programs.

There has been and, probably always will be, what we might call traditional cheating -- that is, one student copying from a neighbor (with or without permission), comparing notes during breaks, looking ahead to a separately timed section, etc. New wrinkles have been added with the advent of wireless communications among calculators, electronic notepads, etc.

Students like to be helpful to their friends -- so if one or more students has "looked ahead," they will frequently pass the word during a break, or engage in a group discussion of the topic. Of course, more resourceful and planful students will have planted textbooks, notes, etc. in the restroom for use as needed.

With the globalization of the economy and the proliferation of families being moved around the world by multinational employers, many students now have friends who live and attend school in many parts of the world. With the ease of international communications, students are able to talk regularly with their friends who happen to be living on other continents. It is no great step for these students to take advantage of the time-zone differences for tests that are given internationally. Constructed-response questions seem to lend themselves easily to being woven into an international conversation. For example, there have been cases where an American student in Singapore or Israel has called a friend in the U.S. to discuss the free-response questions a few hours before the U.S. student is to take the test. Knowing the topic(s) in time to review one's text or notes can be a great help.

Time-zone differences aren't limited to calls among friends or discussion of document-based questions. One particularly helpful coaching operation had a knowledgeable confederate take the test in an earlier time zone and transmit the multiple-choice answers which were then encoded on #2 pencils and handed to the participating test-takers as they entered the test center on the west coast.

Of course, there is old fashion theft of test books to gain advance knowledge of the test question-- where a student either breaks into a secure storage area -- would you believe crawling through the ventilating ducts to get into a storage area?? -- Or takes advantage of sloppy handling by the test administrator. Sometimes what appears to be "sloppy" procedures is really a reflection of the common phenomenon of the left hand not knowing what the right hand is doing. For example, a test book was stolen from a locked store room to which only the principal and the test administrator had a key -- or so they thought. Investigation turned up the fact that a night school was also held in that building. The administration of the night school also had access to that store room and left it open, not knowing that secure exams were being stored there. Of course, there are many cases where custodians or other service personnel have access to storage areas for which the test administrator believes he/she is the sole key-holder.

Teachers and other educators also play a role in undermining the validity of the tests. At one level, teachers choose to ignore evidence that a student or students have pre-knowledge of the content of the test. "These are all highly motivated, morally upstanding young people who wouldn't think of cheating." In one of the cases using time-zone differences to gain advance knowledge, the student came to school and teased the teacher about having made a wrong prediction about the topic of a major free-response question. The teacher asked the student how she knew and the student told her she had talked with her friend overseas. To which the teacher replied, "I have to go move my car," leaving the student to share the information with her classmates prior to the beginning of the testing period.

In other cases, educators tacitly engage in collusion with the students. This may take the form of leaving the administration of a high-stakes test

unsupervised for a period of time or answering questions and giving hints during a break in the test administration -- even when they are not the proctor. In Advance Placement, it is not uncommon for the teacher (who can not supervise the administration of his/her own subject) to provide donuts, etc. during the break as an encouragement to the students. Students who have gained advance knowledge by opening part II before the break, can pose some useful questions to the teacher.

Another way of tacitly supporting students is to leave useful information on display in the testing room. This may take the form of review session notes being left on the chalk board, or time-lines, charts or other visuals being left on the walls.

Teachers want their students to do well on external tests and will focus their teaching attention on topics or skills that they believe will be tested. We are all familiar with the well-documented phenomenon of within-school (or district) scores increasing over time as the same form of a standardized test is used repeatedly. However, this same phenomenon can happen with alternate forms that use common items for equating purposes. Take, for example, the teacher who carefully organizes his/her class so that every student is assigned 3 or 4 m/c questions to be remembered (preferably copied) and brought out of the testing session, ostensibly for review and re-teaching. If the teacher uses those questions with future classes, the next time a block of those questions is used as embedded equating items for a new form, students will be pleasantly surprised to encounter a number of familiar questions.

Of course, teachers can be more deliberate and systematic in their collection of supposedly-secure test questions. There was the teacher who was also a long-term test supervisor for one of the testing programs. He always used that opportunity to go to the photocopying machine with a test book in hand. He also ran a coaching (tutoring) program on the side. Because there was only a modest number of different forms of each subject being used in administrations throughout the year, he became quite successful at preparing students for the tests.

SO WHAT CAN WE DO?

An extended dialogue around this topic is needed within the measurement and education communities. Let me make a few suggestions, recognizing that virtually every possibility has serious financial implications for the testing program or schools.

1) We in the measurement community can do a better job of educating / persuading test administrators, teachers, parents, students of the importance of standard conditions for certain kinds of testing purposes. There is a great deal of emphasis in the educational system on adapting the institution(s) to meet individual needs--and I suspect that most of us would agree with that emphasis for many aspects of schooling. However, we need to help others stand that testing and assessment used for high stakes, comparative purposes depend on a standard measurement process. Most of them would not want to buy a piece of property that was described in idiosyncratic units of measurement. Similarly, a test's description of academic achievement needs to be expressed in terms that have a common meaning for all test-takers. We need to make clear that it is both unfair to other test takers and that the validity of a test is undermined if some students have advance knowledge of the questions or receive help in responding to the questions. I've become convinced that many educators do not understand the purpose of standardization nor the elements which can affect the interpretation and use of the results. At ETS, we have recently begun a series of workshops for test center administrators that bring them back to some of these basic ideas and reinforce the fact that the best way to avoid security problems is to meticulously follow the instructions provided with the test. This seems to be reducing the number of security cases at sites where the test administrators have been through the workshop.

2) When a student has advance knowledge of test questions, it usually becomes known to their peers and frequently to their teachers. Somehow, kids can't keep their illicit knowledge to themselves but seem compelled to discuss it with others. Consequently, many test security cases come to light because someone -- teacher, student, parent -- has the moral courage to say that "This is not fair, it's not right" and notifies the testing agency. I believe

that testing agencies could do more to encourage such behavior and could increase the avenues by which students, parents, or teachers could report situations of advance knowledge or of assistance given or received. Again, this idea is premised on helping all participants understand why there are basic ground rules to provide a standardized situation.

3) It can be useful for the testing agency or program to make some pre-administration audits -- both to check on how tests are received, handled, and stored -- but also to provide some personal professional development for those responsible for administering the test.

4) Several of the potential sources of cheating can be addressed by having multiple forms of the test used at the same administration and/or by fancy packaging that makes it more difficult to look ahead or to return to "completed" sections. Of course, such strategies have major cost implications.

5) Computer-Based Testing, and especially Computer-Adaptive Testing, has the potential of reducing or eliminating some of these threats to the integrity of the test scores. Early experience with CBT for GRE suggests that there is a reduction in security cases. However, like all solutions, CBT undoubtedly contains the seeds of new ways of gaining unfair advantage..

6) The measurement community can strengthen its efforts to educate users that multiple sources of information should be used in making major decisions rather than relying solely on the results of a test. At the same time, it is important for the measurement community to develop additional practical measures or indicators that will help broaden the scope of information available to decision makers.

1. E. F. Lindquist (Ed.) *Educational Measurement*, American Council on Education, Washington DC, 1951.